

# Complexity versus Agreement for Many Views

## Co-regularization for Multi-view Semi-supervised Learning

Odalric-Ambrym Maillard\* and Nicolas Vayatis\*\*

<sup>1</sup> Sequential Learning Project,  
INRIA Lille - Nord Europe, France  
odalric.maillard@inria.fr

<sup>2</sup> ENS Cachan & UniverSud - CMLA UMR CNRS 8536  
nicolas.vayatis@cmla.ens-cachan.fr

**Abstract.** The paper considers the problem of semi-supervised multi-view classification, where each view corresponds to a Reproducing Kernel Hilbert Space. An algorithm based on co-regularization methods with extra penalty terms reflecting smoothness and general agreement properties is proposed. We first provide explicit tight control on the Rademacher ( $L_1$ ) complexity of the corresponding class of learners for arbitrary many views, then give the asymptotic behavior of the bounds when the co-regularization term increases, making explicit the relation between *consistency* of the views and *reduction* of the search space. Since many views involve many parameters, we third provide a parameter selection procedure, based on the stability approach with clustering and localization arguments. To this aim, we give an explicit bound on the variance ( $L_2$ -diameter) of the class of functions. Finally we illustrate the algorithm through simulations on toy examples.

## 1 Introduction

In real-life applications for classification tasks, different representations of a same object may be available. Financial experts may use different sets of indicators to assess the current market regime, while in the context of active computer vision, several views of the same object are provided before rendering the decision. This problem is known as that of multi-view classification. After the early work of (Blum & Mitchell, 1998) on learning from both labeled and unlabeled data, this topic has been considered more recently by several authors (see for example (Sridharan & Kakade, 2008), (Weston et al., 2005), (Zhou et al., 2004)). In (Balcan & Blum, 2005), the authors propose a theoretical PAC-model for semi-supervised learning where multi-view learning appears as a special case. Due to the restriction over the search space (compatibility between different views), multi-view learning may provide good generalization results, and indeed this is the case in numerical experiments (e.g. (Belkin et al., 2005)). In (Rosenberg & Bartlett, 2007), these results are applied to a two-view learning problem

---

\* The first author is eligible for the E.M.Gold Award.

\*\* The second author was partly supported by the ANR Project TAMIS.

and explicit bounds on the Rademacher complexity of the class of predictors are computed. Various algorithms are introduced together with theoretical studies are provided in (Sindhwani et al., 2005). In the latter references, the central issue addressed was to explain how consistency between views affects the performance of classification procedures. Indeed, in multi-view learning, we consider individual predictors based on separate views, and one intuitive idea is that (1) having a good final predictor is related to the agreement of individual predictors on a majority of labels. It is generally assumed that (2) each view is independent from the others conditionally on labeled data. Though this may be weakened (see (Balcan et al., 2005)), providing theoretical justification to the heuristics that conditional independence of the views allows for high-performance results (two compatible classifiers trained on independent views are unlikely to agree on a mislabeled item) has been the motivation for most of the works on this topic. Thus we build on the same heuristics (1) and (2). As we intend to exploit all the information available in the classification task, our setup will also take unlabeled data into consideration.

In the present paper, we consider semi-supervised multi-view binary classification with many views. Allowing for more than two views brings up new questions. For instance, (i) how does the number of views  $V$  affects complexity measures? and (ii) how to choose the parameters when there are as many as  $O(V^2)$  of them? For the first issue, we will focus on the Rademacher average and track down the dependency on  $V$  and other parameters in this formulation. As far as the second issue is concerned, various strategies can be invoked. In supervised classification, cross-validation (e.g. 10-fold) techniques are widely used due to their ease of implementation, but theory is unavailable in most cases (see (Celisse, 2008) and references therein for some recent developments). Another idea comes from recent work on clustering and makes use of the stability approach (see (Ben-David et al., 2006)). This relies on strongly theoretically founded results known as localization arguments (see (Koltchinskii, 2006)) which takes advantage of the so-called ‘small ball estimates’ (see also (Li & Linde, 1999), (Berthet & Shi, 2001)). The stability approach has also been applied successfully to other learning problems (see e.g. (Sindhwani & Rosenberg, 2008)). This is the one we have chosen in order to perform the selection of parameters. Comparison of different selection procedures, although interesting by itself, is not the purpose of this paper.

In the sequel, we introduce an algorithm which combines semi-supervised and multi-view ideas and generalizes over previous algorithms: it contains RLS, co-RLS and co-Laplacian (see (Rosenberg & Bartlett, 2007), (Sindhwani et al., 2005)) algorithms as special cases. We use the setup of Reproducing Kernel Hilbert Spaces (RKHS) and provide explicit upper and lower data-dependent bounds on the Rademacher complexity of the class of functions involved by this general algorithm. Our second contribution is to give a new parameter selection procedure, based on the work of (Koltchinskii, 2006), together with explicit stability bounds ( $L_2$ -diameter) on the localized class for the general algorithm, which has not been investigated so far.

The paper is organized as follows: Section 2 defines our framework and the objective function. Section 3 is devoted to the Rademacher complexity control with the first main theorem (section 3.2), and the asymptotic behavior of the bound. Section 4 presents our stability-based selection procedure and the second main theorem, on the  $L_2$  local diameter of our class of functions. In Section 5, we successfully apply the algorithm to some toy examples.

## 2 Setup for Multi-view Semi-supervised Learning

Our approach is based on penalized empirical risk minimization in RKHS. A compound penalty term will reflect both the complexity of the class of decision functions and the particular context of multi-view semi-supervised learning. This is an important improvement on the work of (Sindhwani et al., 2005) where penalties (and corresponding algorithms) are considered separately. The goal we pursue here is of unifying algorithms instead of comparing them. In this section, we provide the notations and definitions of the penalty terms involved.

In the multi-view setup, an observation results from elements taken in a collection of representation spaces  $\mathcal{X}^{(v)}, v = 1, \dots, V$ , where  $V$  is the number of views. We write  $x = (x^{(1)}, \dots, x^{(V)})$ , where  $x^{(v)} \in \mathcal{X}^{(v)}$ , for the resulting point living in the product space which accounts for the multiple views of the object.

**Learning with RKHS.** Let  $\{-1, 1\}$  be the label set and a loss function, for instance  $\text{loss}_{\text{square}}(g(x), y) = (y - g(x))^2$ , with  $x$  a data point and  $y$  a label. As usual, we consider the label set  $\mathcal{Y} = [-1, 1]$  instead. We consider  $n = u + l$  i.i.d. data points  $l$  of which are labeled, and  $u$  are unlabeled. We now define the loss of a multi-view classifier  $f$  thanks to the corresponding  $f^{(v)}$  in each view:

**Definition 1 (Loss).** For  $f = (f^{(1)}, \dots, f^{(V)})$  and a sample  $(x_i, y_i)_{i=1, \dots, l}$ :

$$\text{Loss}(f) = \frac{1}{l} \sum_{i=1}^l \text{loss}(f, x_i, y_i),$$

where  $\text{loss}(f, x, y)$  may be for instance  $\frac{1}{V} \sum_{v=1}^V \text{loss}_{\text{square}}(f^{(v)}(x^{(v)}), y)$  or  $\text{loss}_{\text{square}}(\frac{1}{V} \sum_{v=1}^V f^{(v)}(x^{(v)}), y)$ .

We consider real-valued decision functions  $\phi : x \mapsto \frac{1}{V} \sum_{v=1}^V f^{(v)}(x^{(v)})$  where  $f^{(v)} : \mathcal{X}^{(v)} \rightarrow \mathcal{Y}$  is a classifier. We assume that each predictor  $f^{(v)}$  lives in an RKHS  $\mathcal{F}^{(v)}$  with kernel  $K^{(v)}$ , associated representation function  $k_v(\cdot, \cdot)$ , and norm  $\|\cdot\|_v$ . Thanks to the representer theorem, we restrict only to functions  $f^{(v)} \in \mathcal{L}_v = \text{span}\{k_v(x_i^{(v)}, \cdot)\}_{i=1}^{l+u} \subset \mathcal{F}^{(v)}$ . Let  $\mathcal{F}$  be the product space of the views  $\mathcal{F}^{(v)}$  and  $\mathcal{L} \subset \mathcal{F}$  the product space of the spans. This complexity penalization leads to the following definition:

**Definition 2 (Complexity).** For  $f = (f^{(1)}, \dots, f^{(V)}) \in \mathcal{F}$ , we define:

$$\text{Complexity}(f) = \sum_{v=1}^V \lambda_v \|f^{(v)}\|_v^2 \quad \text{where } \lambda \in \mathbb{R}_+^V$$

**Semi-supervised regularization.** In the sequel, we consider a batch of  $n = u+l$  i.i.d. data points,  $(x_i)_{i=1..l,l+1..l+u}$ .  $x_i$  is the representation of one object in all views  $x_i = (x_i^{(1)}, \dots, x_i^{(V)})$ . This setup is in-between classification and clustering theory: the labeled part allows for an objective function (whereas in clustering, there is no labeling, thus no objective truth), and the unlabeled part involves structure detection in the data. Using a graph-Laplacian is a natural choice to express the search for structure, as explained for instance in (Smola & Kondor, 2003; Ando & Zhang, 2007). The idea is to consider that the data points depict a manifold (see (Belkin et al., 2005)), for which the graph-Laplacian is a discretized Laplace-Beltrami differential operator. Assuming we have for each view  $v$  a similarity graph given by its adjacency matrix  $W^{(v)}$ , then the (unnormalized) graph-Laplacian is  $L^{(v)} = D^{(v)} - W^{(v)}$ , where  $D^{(v)}$  is the diagonal matrix  $D_{i,i}^{(v)} = \sum_j W_{i,j}^{(v)}$ . Other interesting choices are the symmetrical or random walk normalized graph-Laplacian. Since intuitively one wants that each  $f^{(v)} \in \mathcal{F}^v$  be smooth w.r.t similarity structures in all views, we use the weighted average graph-Laplacian  $L = \sum_{v=1}^V \alpha_v L^{(v)}$  with weights  $\alpha$  summing to 1.

**Definition 3 (Smoothness).** For  $f = (f^{(1)}, \dots, f^{(V)})$ , we define:

$$\text{Smoothness}(f) = \sum_{v=1}^V \gamma_v \mathbf{f}^{(v)T} L \mathbf{f}^{(v)}, \text{ where}$$

- $\gamma = (\gamma_1, \dots, \gamma_V) \geq 0$  meaning that each component is positive.
- $L$  is defined based on  $L^{(v)}$ , the graph-Laplacian corresponding to the  $v$ -th view:  $L = \sum_{v=1}^V \alpha_v L^{(v)}$  with  $\sum_{v=1}^V \alpha_v = 1$ .
- $\mathbf{f}^{(v)}$  is the vector  $(f^{(v)}(x_1^{(v)}), \dots, f^{(v)}(x_{l+u}^{(v)}))^T$ .

**Multiple view co-regularization.** In a multi-view approach, the need for compatibility between the  $f^{(v)}$  is conveyed by a so-called Agreement term. We propose the following one which penalizes disagreement with a square loss and generalizes (Sindhwani et al., 2005) to our setting:

**Definition 4 (Agreement).** For  $f = (f^{(1)}, \dots, f^{(V)})$ , and symmetric positive definite matrices  $c^L, c^U \in \mathbb{R}^{V \times V}$ , we define  $\text{Agreement}(f)$  as the sum of:

$$C^L(f) = \sum_{v_1 \neq v_2} c_{v_1, v_2}^L \sum_{i=1}^l [f^{(v_1)}(x_i^{(v_1)}) - f^{(v_2)}(x_i^{(v_2)})]^2$$

and

$$C^U(f) = \sum_{v_1 \neq v_2} c_{v_1, v_2}^U \sum_{i=l+1}^{l+u} [f^{(v_1)}(x_i^{(v_1)}) - f^{(v_2)}(x_i^{(v_2)})]^2$$

**Compound complexity penalties.** We finally formulate the objective function in this setup as the result of loss minimization with a compound penalty:

– Compute:

$$[1] \quad f^* = \operatorname{argmin}_{f \in \mathcal{F}} \{ \text{Loss}(f) + \text{Complexity}(f) \\ + \text{Smoothness}(f) + \text{Agreement}(f) \}$$

– Output:  $\phi = \frac{1}{V} \sum_{v=1}^V f^{*(v)}$

We point out that there is a representer theorem for this setting. Indeed, for any fixed  $f^{(2)}, \dots, f^{(V)} \in \Pi_{v=2}^V \mathcal{F}^{(v)}$ ,  $f^{*(1)}$  minimizes a function

$$c_{f^{(2)}, \dots, f^{(V)}}(f(x_1^{(1)}), y_1, \dots, f(x_n^{(1)}), y_n) + g_{f^{(2)}, \dots, f^{(V)}}(\|f\|_1)$$

w.r.t.  $f$ . Thus the representer theorem tells us that  $f^{(1)} \in \mathcal{L}_1$ . Iterating the argument leads to  $f^* \in \mathcal{L}$ . We also refer to (Sindhwani & Rosenberg, 2008) for an alternative construction where one single RKHS combines all the views.

For specific choices of the parameters, we recover the former problems studied in previous papers:

- when  $\gamma$  and  $C$  are 0 we have a Regularized Least Squares (RLS) in RKHS,
- when only  $\gamma = 0$  we have a Co-Regularized Least Squares (co-RLS) problem (see (Sindhwani et al., 2005)),
- when Agreement is diagonal nonzero (i.e.  $c^L$  and  $c^U$  are diagonal), we have a co-Laplacian method (e.g. co-Laplacian RLS, co-Laplacian SVM, see (Sindhwani et al., 2005)); indeed, the  $f^{(v)}$  are decoupled, and thus problem [1] amounts to solving for each  $v$ :

$$f^{(v)*} = \operatorname{argmin}_{f^{(v)} \in \mathcal{F}^{(v)}} \text{Loss}(f^{(v)}) + \lambda_v \|f^{(v)}\|_v^2 + \gamma_v \mathbf{f}^{(v)T} L \mathbf{f}^{(v)} .$$

### 3 Excess Risk Bound

This section is devoted to the control of the Rademacher complexity in our problem.

We need the following assumption from (Rosenberg & Bartlett, 2007), which is satisfied for instance by the square loss ( $\text{Loss}(0, \dots, 0) = \frac{1}{l} \sum_{i=1}^l \frac{1}{V} \sum_{v=1}^V y_i^2 \leq 1$ ):

*Assumption A1:* The loss functional satisfies  $\text{Loss}(0, \dots, 0) \leq 1$  where  $(0, \dots, 0)$  is the multi-predictor with constant output 0.

#### 3.1 Preliminaries

One nice property is that under assumption (A1), the final predictor  $\phi$  belongs to:

$$\mathcal{J} = \left\{ x \rightarrow \frac{1}{V} \sum_{v=1}^V f^{(v)}(x^{(v)}) : (f^{(1)}, \dots, f^{(V)}) \in \mathcal{H} \right\}$$

with  $\mathcal{H}$  being the class of multi-predictors  $f$ , with total penalty bounded by 1:

$$\mathcal{H} = \{f \in \mathcal{L} : \text{Complexity}(f) + \text{Smoothness}(f) + \text{Agreement}(f) \leq 1\} .$$

Excess risk bounds involve the Rademacher complexity of the class  $\mathcal{G}$  of learners. For a sample  $(x_1, \dots, x_n)$ , it is defined as

$$R_n(\mathcal{G}) = \mathbb{E}_\sigma \left[ \sup_{g \in \mathcal{G}} \left| \frac{2}{n} \sum_{i=1}^n \sigma_i g(x_i) \right| \right]$$

where  $(\sigma_i)_{i \leq n}$  are Rademacher i.i.d. random variables ( $\mathcal{P}(\sigma_i = 1) = \mathcal{P}(\sigma_i = -1) = \frac{1}{2}$ ).

The following proposition, adapted from (Rosenberg & Bartlett, 2007) makes use of this data-dependent complexity to derive an upper bound of the excess risk:

**Proposition 1 (Excess risk).** *For any positive loss function  $L$  uniformly  $\beta$ -Lipschitz w.r.t its first variable and upper-bounded by 1, then conditionally on the unlabeled data,  $\forall \delta \in (0, 1)$ , with probability at least  $1 - \delta$  over the labeled points drawn i.i.d, for  $\phi_l^*$  the empirical minimizer of the objective function:*

$$\mathbb{E}(L(\phi_l^*(X), Y)) - \inf_{\phi \in \mathcal{J}} \mathbb{E}(L(\phi(X), Y)) \leq 4\beta R_l(\mathcal{J}) + \frac{2}{\sqrt{l}}(2 + 3\sqrt{\ln(2/\delta)/2})$$

The proof is an easy combination of classical generalization bounds with some arguments from (Rosenberg & Bartlett, 2007) and the following contraction principle: if  $h$  is  $\beta$ -Lipschitz and  $h(0) = 0$ , then  $R_n(h \circ \mathcal{J}) \leq 2\beta R_n(\mathcal{J})$  (see (Ledoux & Talagrand, 1991)), together with the symmetry of  $\mathcal{J}$ .

### 3.2 Explicit Rademacher Complexity Bound

*Block-wise notations.* We use the following notations: for any  $n$ ,  $I_n$  or  $I$  is the identity of  $\mathbb{R}^n$ ,  $0_{u,l}$  the zero matrix of  $\mathbb{R}^{u \times l}$ . For any given  $n_1, n_2$ ,  $A^{(v)} \in \mathbb{R}^{n_1 \times n_2}$ ,  $\bar{A}$  is the block-diagonal matrix with blocks  $A^{(v)}$ ,  $v = 1..V$  (of size  $n_1V \times n_2V$ ), and similarly  $\underline{A}$  the block-row matrix of size  $n_1V \times n_2$ . To multiply block-wise each block  $A^{(v)}$  by the  $v$ -th component of a vector  $\lambda \in \mathbb{R}^v$ , let  $\tilde{\lambda}$  be the block-diagonal matrix of size  $n_1V \times n_1V$  with blocks  $\lambda_v I_{n_1}$ . Since we always multiply the  $v$ -th block with the  $v$ -th component, we drop the index.

*Data.* With the following matrices, we decompose between labeled and unlabeled data:  $K^{(v)} = (k_v(x_i^{(v)}, x_j^{(v)}))_{1 \leq i, j \leq n} = \begin{pmatrix} K_L^{(v)} \\ K_U^{(v)} \end{pmatrix} \in \mathbb{R}^{n \times n}$  and  $\Pi = \begin{pmatrix} I_l \\ 0_{u,l} \end{pmatrix} \in \mathbb{R}^{nV \times l}$

*Agreement.* To compare between views pairwise, we introduce a block-line defined  $\delta \in \mathbb{R}^{nV(V-1) \times nV}$ , with blocks  $(0 \dots 0 I_n 0 \dots 0 -I_n 0 \dots 0)$  with identity matrices at position  $v_1$  and  $v_2 \neq v_1$ . Let also the block-diagonal matrix  $C_{v,w}$  with diagonal blocks  $(c_{v,w}^L)_{i=1..l}$  and  $(c_{v,w}^U)_{i=l+1..l+u}$ , and then  $C \in \mathbb{R}^{nV(V-1) \times nV(V-1)}$  the block-diagonal matrix with blocks  $C_{v,w}$  when  $v, w \in \{1, \dots, V\}$

*Smoothness.* Let  $L_I$  be the diagonal block matrix with all  $V$  blocks equaled to  $L$ . Note that we would have introduced  $\tilde{\alpha} \bar{L}$  instead, if we have used each graph Laplacian and not the average Laplacian  $L$  in the smoothness term.

Thanks to the previous notations, we can now state our first main theorem, which shows an explicit upper and lower data-dependent bound for the Rademacher complexity of the class of functions.

**Theorem 1 (Rademacher complexity bound).** *Under assumption (A1), then*

$$\frac{2}{2^{1/4}} \frac{b}{Vl} \leq R_l(\mathcal{J}) \leq \frac{2b}{Vl}$$

where  $b^2 = \text{tr}(B\tilde{\lambda}^{-1}\Pi\underline{K}_L^T) - \text{tr}(J'^T(I + M')^{-1}J')$  with

- $B = (I + \tilde{\lambda}^{-1}\tilde{\gamma}L_I\overline{K})^{-1} \in \mathbb{R}^{nV \times nV}$
- $J' = \sqrt{C}\delta\tilde{\lambda}^{-1}B^TK_L^T \in \mathbb{R}^{nV(V-1) \times l}$
- $M' = \sqrt{C}\delta\overline{K}B\tilde{\lambda}^{-1}\delta^T\sqrt{C} \in \mathbb{R}^{nV(V-1) \times nV(V-1)}$

Note that  $b$  is explicit as a difference of two terms. The first term only depends on unlabeled data when Smoothness is null, and contains no co-regularization term. The second term corresponds to the idea that there is a reduction in complexity of the space. Indeed, in section 3.3, we give some results about the behavior of  $b$  enforcing this idea. As pointed by (Sindhwani & Rosenberg, 2008), this term is connected to a specific norm induced by the parameters and data over the space.

This Theorem generalizes previous results: for instance, if  $V = 2$ ,  $\gamma = 0$ , and  $c_{v,w}^L = 0$ , we recover exactly the previous known bound of (Rosenberg & Bartlett, 2007) where our  $2c_{v,w}^U$  corresponds to their  $\lambda$  and our  $\lambda$  is their  $(\gamma_{\mathcal{F}}, \gamma_{\mathcal{G}})$ .

### 3.3 Asymptotics

Let  $\theta = (\alpha, \lambda, \gamma, C)$  be the parameters of the learning problem, where  $\alpha$  appears in the graph-Laplacian,  $\lambda$  in the Complexity term,  $\gamma$  in the Smoothness term and  $C$  in the Agreement term. The number of parameters grows with  $O(V^2)$ . We study how the previous Rademacher bound changes with these parameters.

*More agreement reduces space complexity.* The second term appearing in the expression of  $b^2$  depends on the co-regularization (matrix) parameter  $C$ . To see how constrained is the space when using bigger penalization, we introduce  $\Delta(C) = \text{tr}(J'^T(I + M')^{-1}J')$ , which can be written, provided that  $C^{-1}$  exists, as:

$$\Delta(C) = \text{tr}(J_1^T(C^{-1} + M_1)^{-1}J_1)$$

where  $J_1 = \delta\tilde{\lambda}^{-1}B^TK_L^T$  and  $M_1 = \delta\overline{K}B\tilde{\lambda}^{-1}\delta^T$ .

Thus when the eigenvalues of  $C$  increases to  $+\infty$ ,  $\Delta(C)$  tends to:

$$\Delta_\infty = \text{tr}(\underline{K}_L^T B\tilde{\lambda}^{-1}\delta^T(\delta\overline{K}B\tilde{\lambda}^{-1}\delta^T)^{-1}\delta\tilde{\lambda}^{-1}B^TK_L^T),$$

which can be rewritten  $\Delta_\infty = \text{tr}(B\tilde{\lambda}^{-1}\Pi_l\underline{K}_L^T)$ , and shows that  $b^2 \rightarrow 0$  in this case. That  $b$  decreases as the model gets more constraint is coherent with the intuition of multi-view learning. Similarly,  $b^2 \rightarrow 0$  whenever  $\|\gamma\|$ , or  $\|\lambda\| \rightarrow \infty$ .

*Unconstrained space.* When the constraint on the space vanishes, we have a completely different behavior. Indeed, if  $C = 0$  then  $\Delta(C) = 0$ . When  $\gamma = 0$ , we refer to (Rosenberg & Bartlett, 2007). Finally, when  $\lambda = 0$ ,  $b^2$  has the following expression (provided every term appearing in this expression is finite and defined):

$$b^2 = \text{tr}(\Pi_l^T L_I^{-1} \tilde{\gamma}^{-1} \Pi_l) - \text{tr}(\Pi_l^T L_I^{-1} \tilde{\gamma}^{-1} \delta^T (C^{-1} + \delta L_I^{-1} \tilde{\gamma}^{-1} \delta^T)^{-1} \delta L_I^{-1} \tilde{\gamma}^{-1} \Pi_l)$$

Note that when both  $\gamma$  and  $\lambda$  tend to 0, the previous bound may tend to  $\infty$  even in some simple case (which is coherent with the intuition). Note also that the dependency with  $V$  is hidden here in the trace.

## 4 Stability-Based Parameter Selection

The multi-view setting involves new questions, like the choice of the parameters since there are  $O(V^2)$  many of them. We now describe an automatic parameter selection procedure which will be theoretically sound.

### 4.1 Theoretical Selection Procedure

Let  $\mathcal{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  be the empirical measure, and  $P$  the true measure. Thus  $\mathcal{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$  and  $\mathcal{P} f = \mathbb{E}(f(X))$ . For a general class  $\mathcal{F}$  of functions, and probability measure  $Q$ , we define  $\mathcal{F}_Q(\epsilon) = \{f \in \mathcal{F}; Qf - \inf Qf \leq \epsilon\}$  and then introduce the true  $\epsilon$ -optimal ball  $\mathcal{F}(\epsilon) = \mathcal{F}_{\mathcal{P}}(\epsilon)$ , and the empirical  $\epsilon$ -optimal ball  $\mathcal{F}_n(\epsilon) = \mathcal{F}_{\mathcal{P}_n}(\epsilon)$ , or balls around the Empirical Risk Minimizer (ERM) and True Risk Minimizer (TRM). For a general class  $\mathcal{F}$  of functions, we now assume that we have  $T : \mathcal{F}^2 \rightarrow \mathbb{R}^+$  such that  $\forall f, g \in \mathcal{F} \quad \mathbb{V}(f - g) \leq T^2(f, g)$ , and then introduce the two objects:  $\Delta_n(\epsilon) = \sup_{f_1, f_2 \in \mathcal{F}(\epsilon)} |P_n - P|(f_1 - f_2)$  and  $D_{\mathcal{F}}(\epsilon) = \sup_{f, g \in \mathcal{F}(\epsilon)} T(f, g)$ . We refer to the first one as a  $L_1, P$ -diameter and the second one as a  $L_2, P$ -diameter. Lemma 1 in (Koltchinskii, 2006) tells us that for large enough radii, the empirical and true quasi-optimal sets around the ERM and TRM are included in each other, or put differently, that true quasi-optimal sets can be estimated by empirical quasi-optimal sets:

**Lemma 1.** (Koltchinskii) *For any  $\epsilon > 0$ , and any  $\lambda < 1$ , we set*

$$B_n(\epsilon, \lambda) = 2 \frac{\Delta_n(\epsilon)}{\lambda} + \frac{\log(\epsilon^{-1})}{\lambda n} + \frac{2}{\lambda} \sqrt{\frac{2 \log(\epsilon^{-1})}{n} [D_{\mathcal{F}}^2(\epsilon) + 2\Delta_n(\epsilon)]},$$

$$\text{and } r_n(\epsilon, \lambda) = \inf \left\{ \alpha \in [0, 1]; \sup_{j \in \mathbb{Z}; 1 \geq \lambda^j \geq \alpha} B_n(\epsilon, \lambda^j) \leq \lambda \right\} .$$

Set also  $\epsilon' = \left(2 + \frac{\ln(r_n(\epsilon, \lambda))}{\ln(\lambda)}\right) \epsilon$ . Then, with probability larger than  $1 - \epsilon'$ :

$$\forall r \geq r_n(\epsilon, \lambda) \quad \mathcal{F}(r) \subset \mathcal{F}_n(3r/2) \text{ and } \mathcal{F}_n(r) \subset \mathcal{F}(2r) .$$

In the general case, if the radii are too small, then such inclusions no longer hold, and the intersection may even be empty. For our problem, we will simply select the parameter  $\theta$  inducing the larger range of quasi-optimal sets controlled around the ERM, which is a notion of stability. Thus, for a given radius  $\epsilon$  of the true penalized ball, we want to minimize the critical radius  $r_n$  w.r.t.  $\theta$ . A side motivating intuition is that having good stability allows for easy discovery of the minimizer  $f^*$ .

### 4.2 Empirical Selection Procedure

We now propose an empirical version of this lemma. Fortunately, using an empirical estimation of the  $r_n(\epsilon, \lambda)$  is possible thanks to the Theorem 3, page 18, in (Koltchinskii, 2006), leading to a full data-dependent quantity. Indeed, let  $\hat{\Delta}_n(\epsilon) = R_n(\mathcal{F}_n(\epsilon))$  and  $\hat{D}_{\mathcal{F}_n}(\epsilon) = \sup_{f,g \in \mathcal{F}_n(\epsilon)} T_n(f, g)$ , with  $T_n^2$  bounding the empirical variance  $\mathbb{V}_n$ . The empirical versions of  $B_n(\epsilon, \lambda)$  and  $r_n(\epsilon, \lambda)$  given by (Koltchinskii, 2006) are:

$$\hat{r}_n(\epsilon, \lambda) = \inf \left\{ \alpha \in [0, 1]; \sup_{j \in \mathbb{Z}; 1 \geq \lambda^j \geq \alpha} \hat{B}_n(\epsilon, \lambda^j) \leq \lambda^3 \right\}, \text{ where}$$

$$\hat{B}_n(\epsilon, \lambda) = \frac{2c\hat{\Delta}_n(c'\epsilon)}{\lambda} + 2\hat{D}_{\mathcal{F}_n}(c'\epsilon) \sqrt{\frac{\log(\epsilon^{-1})}{\lambda^2 n} + \frac{\log(\epsilon^{-1})}{\lambda n}}$$

and  $c, c' \geq 1$  are universal constants.

We now propose to apply this result to semi-supervised multi-view classification. We identify the classes  $\hat{\mathcal{F}}_{\theta, n}$  to be

$$\mathcal{J}(r) = \left\{ x \rightarrow \frac{1}{V} \sum_{v=1}^V f^{(v)}(x^{(v)}); \quad f \in \mathcal{H}(r) \right\},$$

where  $\mathcal{H}(r) = \{f; \hat{\pi}_{\theta, l}(f) \leq r\}$ , and estimate  $R_n(\hat{\mathcal{F}}_{\theta, n}(\epsilon))$  and  $\hat{D}_{\hat{\mathcal{F}}_{\theta, n}}(\epsilon)$  for each parameter  $\theta$ . Note that the dependency w.r.t.  $\theta = (\alpha, \lambda, \gamma, \mathcal{C})$  is hidden in the definition. Thus we need to bound the Rademacher complexity of  $\mathcal{J}(r)$  and its  $L_2, P_n$ -Diameter. An analysis of the proof of Theorem 1 shows that changing  $\mathcal{J} = \mathcal{J}(1)$  for  $\mathcal{J}(r)$  affects the Rademacher bound with a factor  $\sqrt{r}$ , leading to a bound  $\frac{2b(\theta)\sqrt{r}}{V}$  for the first term. Following the same analysis as for the  $L_1$ -diameter (or Rademacher complexity), the next theorem gives us the second bound we need:

**Theorem 2 (Empirical local  $L_2$  diameter).** *Under assumption A1, then*

$$\hat{D}_{\mathcal{J}}(r) \leq \frac{2d\sqrt{r}}{\sqrt{IV}}$$

where  $d^2$  is the largest eigenvalue of  $(B - J_2^T(I + M)^{-1}J_2)\tilde{\lambda}^{-1}\Pi(\underline{K}_L^T)^T$ , with  $J_2 = \sqrt{C}\delta\tilde{\lambda}^{-1}B^T$

Note the dependency with  $\sqrt{l}$  instead of the  $l$  for the Rademacher bound.

Eventually, each  $\theta$  leads to a radius  $r_n^\theta(\epsilon, \lambda) \geq \hat{r}_n^\theta(\epsilon, \lambda)$  defined likewise, using upper bounds of Theorem 1 and 2. For maximal stability, we propose to have the largest range of values for which Lemma 1 still holds, which boils down to minimizing this quantity with  $\theta$ . This leads to the following selection procedure where each term is computable:

- Fix a probability threshold with  $\epsilon > 0$  and  $\lambda < 1$ .
- Compute  $r(\theta, n, l, \epsilon, \lambda)$ , defined by:
 
$$\inf \left\{ \alpha \in [0, 1]; \sup_{j \in \mathbb{Z}; 1 \geq \lambda^j \geq \alpha} \tilde{B}_{n,l}(\theta, \epsilon, \lambda^j) \leq \lambda^3 \right\},$$

where the term  $\tilde{B}_{n,l}(\epsilon, \lambda)$  is:

$$\frac{2cb(\theta)\sqrt{c'\epsilon}}{lV\lambda} + \frac{4d(\theta)\sqrt{c'\epsilon}}{\sqrt{l}V} \sqrt{\frac{\log(\epsilon^{-1})}{\lambda^2 n}} + \frac{\log(\epsilon^{-1})}{\lambda n}$$

- Output:
 
$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} r(\theta, n, l, \epsilon, \lambda)$$

## 5 Experiments

We have performed some toy simulations to see the flexibility of this general algorithm and the results are promising. Based on only one or two labeled points, we can always recover perfect labeling of the data, even on the challenging cross-moons data set on which all classical algorithms (Co-Laplacian and Co-RLS) performs badly. For completeness, we first give hints how to solve the minimization problem. Recall that the solutions of [1] can be written  $f^{(v)}(x^{(v)}) = \sum_{i=1}^{l+u} \alpha_i^{(v)} K^{(v)}(x^{(v)}, x_i^{(v)}) = K_{x^{(v)}}^{(v)} \alpha^{(v)}$ . We first consider the case where the loss function is differentiable.

**Theorem 3 (Solution in the differentiable case).** *Assuming that the loss function satisfies  $\nabla_{\alpha^{(v)}} \text{Loss}(f^{(v)}) = 2K^{(v)} A^{(v)} \alpha^{(v)}$ , then the solution of the problem 1 is given by the resolution of the linear system, where the  $\alpha^{(v)}$  are the unknown vectors.*

$\forall v \in 1 \dots V:$

$$Y = [A^{(v)} + \lambda_v I + \gamma_v LK^{(v)}] \alpha^{(v)} + 2 \sum_{w=1}^V C_{v,w} (K^{(v)} \alpha^{(v)} - K^{(w)} \alpha^{(w)})$$

where  $Y_i = y_i$  for  $1 \leq i \leq l$  and  $Y_i = 0$  for  $l + 1 \leq i \leq l + u$ .

The proof is a straightforward application of usual algebra and is omitted here. This system contains as a special case the linear system of (Sindhwani et al.,

2005). We can rewrite it as  $S\alpha = \tilde{Y}$  where  $S$  is an appropriate matrix,  $\alpha = (\alpha^{(1)T}, \dots, \alpha^{(V)T})^T$  and  $\tilde{Y} = (Y^T, \dots, Y^T)^T$ , but  $S$  *a priori* is not positive and may have a very large conditioning number.

An important case of non-differentiable loss function (which is not covered by the previous Theorem) is the hinge loss used in SVM. How to use a classical SVM solver for our problem, is left aside in this paper. A complete derivation is given in (Belkin et al., 2005) when  $\gamma = 0$ .

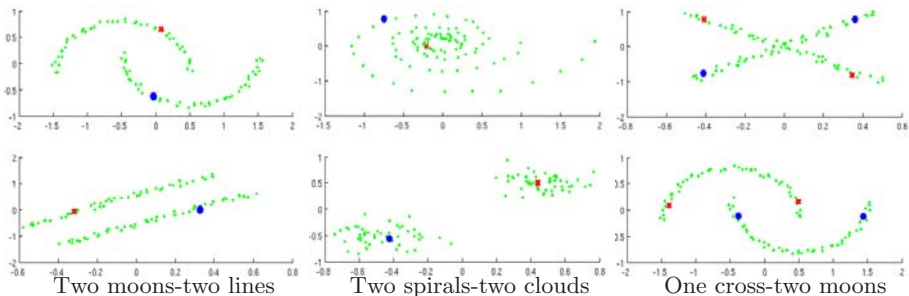
## 5.1 Toy Examples

We have done some experiments on three toy examples (Figure 1), with only two views and two classes for simplicity.

- The easy two moons-two lines data set, for which the data is linearly separable in the second view, and almost separated in the first.
- The more complex two spirals-two clouds data set, with intricate spirals (to “force” the use of graph-Laplacian). Note that a human operator cannot separate the two classes without the information of the second view.
- The challenging cross-two moons data set, which appears to fool the tested algorithms based on only one of the Smoothness or Agreement term.

Since the less labeled object, the more heuristic the definition of the “true” classes, we refer here to human beings to say what are the true classes. Such a definition of truth is a real problem still unsolved in the clustering community and we do not pretend here to solve it. In the first two data sets, a human only needs one label object of each class to recover the classes. For the last one, because the cross yields ambiguity, a human operator needs two objects in each class. Thus, we use this number of labels.

For each algorithm we use the quadratic loss, which is differentiable. The first one is the classical RLS, for which Smoothness and Agreement are set to 0. The second one is a co-RLS, with only Smoothness set to 0. Then we used



**Fig. 1.** Three toy data sets. Normal points for unlabeled points, circle for class number one and cross for class number two. From left to right: Two moons (above)- two lines (below), with one labeled object in each class. Two spirals-two clouds, with one labeled object in each class. One cross-two moons, with two labeled objects in each class.

a Laplacian-based algorithm (co-Laplacian), which outperform co-RLS on the tricky two spirals-two clouds data set, and finally an algorithm with none of the terms set to 0. Since all these algorithms are specialization of the general algorithm, with some parameters set to 0 to highlight some behaviors, we just tuned the parameters by hand trying to find the best results for each algorithm.

Finally, note that the choice of the kernels for each view is important, and we used well-suited kernels for each problem (gaussian for clouds, linear for lines, ...).

algo	dataset 1	dataset 2	dataset 3
RLS	$0.455 \pm 0.035$	$0.103 \pm 0.024$	$0.379 \pm 0.026$
co-RLS	$0.146 \pm 0.071$	$0.103 \pm 0.024$	$0.467 \pm 0.025$
co-Laplacian	$0.242 \pm 0.040$	$0.001 \pm 0.004$	$0.510 \pm 0.028$
general	$0.011 \pm 0.015$	$0.322 \pm 0.067$	$0.042 \pm 0.071$

Empirical misclassification errors for the above algorithms (one set of parameters per dataset, some possibly put to zero when specified to each algorithm), averaged over 1000 runs.

## 6 Discussion and Conclusion

In this paper, we have combined different aspects of semi-supervised and multi-view learning into one algorithm. Based on previous work, we have derived an explicit control for the  $L_1$ -diameter (Rademacher complexity) of the class of decision functions for this new algorithm. Besides, we have shown how considering the full multi-view learning problem may generate new questions. Combining stability ideas from the statistical and clustering community, we have proposed a new stability-based parameter selection procedure, which benefits from strong recent theoretical developments. For this procedure to be implementable, we have controlled the  $L_2$ -diameter of the class as well, which has not been investigated so far for similar settings.

## References

- Ando, R.K., Zhang, T.: Learning on graph with laplacian regularization. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in neural information processing systems, vol. 19, pp. 25–32. MIT Press, Cambridge (2007)
- Balcan, M., Blum, A.: A PAC-style model for learning from labeled and unlabeled data. In: Auer, P., Meir, R. (eds.) COLT 2005. LNCS (LNAI), vol. 3559, pp. 111–126. Springer, Heidelberg (2005)
- Balcan, M.F., Blum, A., Yang, K.: Co-training and expansion: Towards bridging theory and practice. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in neural information processing systems, vol. 17, pp. 89–96. MIT Press, Cambridge (2005)
- Belkin, M., Niyogi, P., Sindhvani, V.: On Manifold Regularization. In: AISTAT (2005)

- Ben-David, S., von Luxburg, U., Pal, D.: A sober look at clustering stability. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, pp. 5–19. Springer, Heidelberg (2006)
- Berthet, P., Shi, Z.: Small ball estimates for brownian motion under a weighted sup-norm. *Studia Sci. Math. Hung.* 1–2, 275–289 (2001)
- Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT 1998: Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100. ACM, New York (1998)
- Celisse, A.: Model selection via cross-validation in density estimation, regression, and change-points detection. Doctoral dissertation, Universite Paris Sud, Faculte des Sciences d’Orsay (2008)
- Golub, G.H., Van Loan, C.F.: Matrix computations. The Johns Hopkins University Press (1996)
- Koltchinskii, V.: Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34(6), 2593–2656 (2006)
- Ledoux, M., Talagrand, M.: Probability on banach spaces: Isoperimetry and processes. Springer, Berlin (1991)
- Li, W.V., Linde, W.: Approximation, metric entropy and small ball estimates for gaussian measures. *Ann. Probab.* 27, 1556–1578 (1999)
- Rosenberg, D., Bartlett, P.L.: The rademacher complexity of co-regularized kernel classes. In: Proceedings of the Eleventh ICAIS (2007)
- Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semi-supervised learning with multiple views. In: Workshop on Learning with Multiple Views, Proceedings of International Conference on Machine Learning (2005)
- Sindhwani, V., Rosenberg, D.S.: An rkhs for multi-view learning and manifold co-regularization. In: ICML 2008: Proceedings of the 25th international conference on Machine learning, pp. 976–983. ACM, New York (2008)
- Smola, A.J., Kondor, R.: Kernels and regularization on graphs. In: Conference on Learning Theory and 7th Kernel Workshop, pp. 144–158 (2003)
- Sridharan, K., Kakade, S.M.: An information theoretic framework for multi-view learning. In: COLT, pp. 403–414. Omnipress (2008)
- Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., Noble, W.S.: Semi-supervised protein classification using cluster kernels. *Bioinformatics* 21, 3241–3247 (2005)
- Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, vol. 16, pp. 321–328. MIT Press, Cambridge (2004)

## Appendix - Proofs

### Sketch of Proof of Theorem 1

The proof of Theorem 1 follows the same line as (Rosenberg & Bartlett, 2007) and extends their result to the compound regularization penalty in the case of an arbitrary number of views. Since there is no novelty in the proof technique, we do not reproduce it here entirely. For completeness, we recall the main next steps: (i) use classical invariance properties of the kernel function to reformulate the optimization problem with an invertible matrix, (ii) apply Lemma 3 below to get the solution, (iii) eventually, rewrite it with the formulation involving the initial data by use of the Sherman-Morrison-Woodbury formula (Golub & Van Loan, 1996). We provide the key intermediate steps adapted to our setting.

**Lemma 2.** *Under assumption (A1), the solution of the minimization problem [1] belongs to the set  $\mathcal{L} \cap \mathcal{H}$ .*

*Proof:* Let  $Q$  be the functional to be minimized, decomposed as:  $Q(f) = \text{Loss}(f) + \Pi(f)$ . For the null multi-view predictor  $0 \in \mathcal{F}$ , we have  $Q(0) = \text{Loss}(0)$ , thus under assumption (A1),  $\inf Q \leq 1$ . But since all terms of  $Q$  are non negative, the solution is in  $\mathcal{H}$ . Finally, that  $f^* \in \mathcal{L}$  by the representer theorem.  $\square$

First, we apply Lemma 2 to reduce the search space. Then, if  $f \in \mathcal{L} \cap \mathcal{H}$ , thanks to the representer theorem, we can write its component in each view  $f^{(v)} = f_{\alpha^{(v)}}^{(v)} = \sum_{i=1}^n \alpha_i^{(v)} k_v(\cdot, x_i^{(v)})$ , where  $\alpha^{(v)} \in \mathbb{R}^n$ . Thus, a matrix reformulation of  $f \in \mathcal{L} \cap \mathcal{H}$  is:

$$f \in \{(f_{\alpha^{(1)}}, \dots, f_{\alpha^{(V)}}) : \underline{\alpha}^T N \underline{\alpha} \leq 1\}$$

where  $\alpha \in \mathbb{R}^{n \times V}$ , and the data-dependent  $N$  square matrix is<sup>1</sup>:

$$N = \tilde{\lambda} \bar{K} + \tilde{\gamma} \text{Diag}(K^{(1)} L K^{(1)} \dots K^{(V)} L K^{(V)}) + \sum_{v_1 \neq v_2} K_C^{v_1, v_2}, \text{ and}$$

$$K_C^{v_1, v_2} = \begin{pmatrix} 0 \\ \vdots \\ K^{(v_1)} \\ \vdots \\ -K^{(v_2)} \\ \vdots \\ 0 \end{pmatrix} C_{v_1, v_2} \begin{pmatrix} 0 \\ \vdots \\ K^{(v_1)} \\ \vdots \\ -K^{(v_2)} \\ \vdots \\ 0 \end{pmatrix}^T.$$

Thus, the definition of the Rademacher complexity can be seen as the solution to an optimization problem under quadratic constraint. Indeed, since  $\mathcal{H}$  is symmetrical:

$$R_l(\mathcal{J}) = \frac{2}{lV} \mathbb{E}_\sigma \sup_{\alpha: \alpha^T N \alpha \leq 1} \alpha^T \underline{K}_L^T \sigma$$

with  $\sigma = (\sigma_1, \dots, \sigma_l)^T \in \mathbb{R}^{l \times 1}$ . To apply Lemma 3, we need an invertible matrix. Let  $P, \Sigma$  such that  $P^{(v)T} K^{(v)} P^{(v)} = \Sigma^{(v)}$  is the diagonal matrix of non zero eigenvalues of  $K^{(v)}$ . We introduce  $\alpha_{//}^{(v)}$ , the projection of  $\alpha^{(v)}$  on the subspace associated to the rows of  $K^{(v)}$ . Since  $\underline{\alpha}^T N \underline{\alpha}$  is left unchanged under this projection, we rewrite it with  $a^{(v)}$  such that  $P^{(v)} a^{(v)} = \alpha_{//}^{(v)}$ , ending up with the constraint  $\underline{a}^T T \underline{a} \leq 1$  where  $T$  is now an invertible matrix. As mentioned, we use the following lemmas to conclude:

<sup>1</sup>  $\text{Diag}(v_1 \dots v_k)$  is a shortcut notation for the square matrix with diagonal blocks  $v_1, \dots, v_k$  on the diagonal.

**Lemma 3.** *If  $M$  is a symmetric positive definite matrix, then*

$$\sup_{\alpha: \alpha^T M \alpha \leq 1} v^T \alpha = \|M^{-1/2} v\| .$$

**Lemma 4.** *(Sherman-Morrison-Woodbury formula) Provided that the inverses exist:*

$$(A + UU^T)^{-1} = A^{-1} - A^{-1}U(I + U^T A^{-1}U)^{-1}U^T A^{-1} .$$

**Sketch of Proof of Theorem 2**

The proof essentially follows the same steps as Theorem 1, but uses Lemma 5 below to solve the minimization problem.

By definition, we have  $\hat{D}_{\mathcal{J}(r)} = \sup_{\phi_1, \phi_2 \in \mathcal{J}(r)} \mathbb{V}_l(\phi_1 - \phi_2)^{1/2}$ , which is also

$$\sup_{\phi_1, \phi_2 \in \mathcal{J}(r)} (\mathcal{P}_l((\phi_1 - \phi_2)^2))^{1/2} \leq \left[ \frac{4}{l} \sup_{\phi \in \mathcal{J}(r)} \sum_{i=1}^l \phi(x_i)^2 \right]^{1/2} .$$

Since  $f^{(v)}(x_i) = K_i^{(v)} \alpha^{(v)}$ , where  $K_i^{(v)}$  is the  $i$ th row of  $K^{(v)}$ ,  $\hat{D}_{\mathcal{J}(r)}^2 \leq \frac{4}{\sqrt{2}l} \underline{\alpha}^T D \underline{\alpha}$  where  $D \in \mathbb{R}^{n^V \times n^V}$  is the symmetrical matrix with block  $v, w$  equal to  $\sum_{i=1}^l K_i^{(v)T} K_i^{(w)}$ . Applying Lemma 2 with (A1),  $\phi \in \mathcal{J}(r)$  is  $\underline{\alpha}^T N \underline{\alpha} \leq r$ . If we introduce the same transformation as in Theorem 1, this is again  $\underline{a}^T T \underline{a} \leq r$  with now invertible  $T$ . Moreover  $T$  as the appropriate form  $A + UU^T$  for Lemma 4. Thus Lemma 5 first tells us that we want the highest eigenvalue of  $T^{-1} \overline{P}^T D \overline{P}$ .

**Lemma 5.** *When  $M$  is symmetric positive definite and  $Q$  is symmetric positive semidefinite, the quadratic problem :*

$$\sup_{a: a^T M a \leq r} a^T Q a$$

*admits as solution  $\lambda r$ , where  $\lambda$  is the highest eigenvalue of  $M^{-1}Q$ .*

Now,  $D = \underline{K} e \underline{K}^T$  with  $e \in \mathbb{R}^{n \times n}$  being the projection matrix with diagonal blocks  $I_l$  and  $0_u$ . Lemma 4 applies to  $T^{-1}$ , and since the eigenvalues of  $T^{-1} \overline{P}^T D \overline{P}$  and  $\overline{P} T^{-1} \overline{P}^T D$  are the same, we compute:

$$\overline{P} A^{-1} \overline{P}^T D = \overline{P} \overline{P}^T B \overline{P} \tilde{\lambda}^{-1} \overline{\Sigma}^{-1} \overline{P}^T \underline{K} e \underline{K}^T = B \tilde{\lambda}^{-1} \Pi (\underline{K}_L^T)^T$$

Where  $A$  comes from Lemma 4. Similar computations yield the second term and allow to conclude the proof.